Enhancing cell culture understanding in the development of biopharmaceuticals by integrated first-principle modelling and machine-learning

Edoardo Tamiazzo,a Gianmarco Barberi,a Paloma Diaz-Fernandez,b Gary Finka,b Pierantonio Facco a,\*

*aCAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo, 9 – 35131 Padova PD, Italy*

*bBiopharm Process Research, Drug Substance Development, GSK, Gunnels Wood Rd, SG1 2NY Stevenage, UK*

\*pierantonio.facco@unipd.it

Abstract

Monoclonal antibodies (mAbs) are biopharmaceuticals which are used to treat a variety of diseases, including cancer and autoimmune disorders. These products are typically produced in mammalian Chinese Hamster Ovary (CHO) cell cultures whose complex and variable nature poses a significant challenge to product and process development.
This study aims at understanding the phenomena occurring in mammalian cell cultures and relating the dynamics of cell metabolism to the macroscopic chemical, physical and biological phenomena occurring in the cell cultures through first-principle modelling and machine learning. In particular, we propose a novel approach in which an improved cell kinetic model is used to describe the macroscopic behavior of the cell cultures. Then, metabolomics data, which provide information on the metabolic dynamics of the cells along the culture, are used to relate the cell metabolism to the phenomena described by the kinetic model through supervised machine learning. The approach is applied to an industrial case study for the mAbs development in AMBR15® miniature bioreactors. A strong relationship of chemical-physical and biological phenomena with the dynamics of cell metabolism is found (average determination coefficients in cross-validation in the range of ~46-90%). Furthermore, the proposed approach allows understanding the effect of cell metabolism on the macroscopic phenomena occurring in the system.

**Keywords**: monoclonal antibody, mechanistic modelling, parameter estimation, metabolomics, machine learning.

* 1. Introduction

Biopharmaceuticals are drugs and treatments derived from genetically engineered living organisms and designed to produce specific biological substances with therapeutic properties. At the forefront of the biopharmaceutical market monoclonal antibodies (mAbs) (Walsh & Walsh, 2022) have gained significant popularity due to their critical role in treating a wide range of diseases. Monoclonal antibodies are typically cultured in CHO mammalian cells in fed-batch bioreactors which contain media with all the essential macro- and micronutrients required for cells optimal growth and viability. Fed-batch cultures are also provided with daily boluses of nutrients, containing the primary sources of carbon such as Glucose, Glutamate and Glutamine, and amino acids.

The characterization of this type of system requires the use of all the information obtained from the analysis of the cell cultures (Barberi et al., 2022). The types of information which may be available in CHO cell culture are: *i)* process data, which describe the macroscopic behavior of cell cultures, and *ii)* metabolomics data, which provide valuable insights into the microscopic metabolic reactions occurring in the system.

State-of-the-art first-principle models for CHO cell cultures rely on kinetic models (Kyriakopoulos et al., 2018), whose parameters are estimated from process data by fitting experimental data. Model parameters embed a strong physical meaning as they are associated with chemical, physical and biological phenomena occurring in the system (Kontoravdi et al., 2010). However, this model does not take into account the vast amount of information on cell metabolism. This is mainly due to the challenge of dealing with the large number of metabolites measured as ions during an experimental batch.

Some models attempts to integrate -omics data in model’s equations adding metabolites as intermediates in equilibrium and using them to fit new parameters (Ahn and Antoniewicz, 2012). However, this increases model complexity and makes the parameters estimation a challenge. Overcoming these issues and exploiting the potential of metabolomics information promises to unlock new knowledge and optimize mAb development and production.

The aim of this work is to integrate the metabolic information on CHO cell cultures (obtained from metabolomics data) and the chemical, physical and biological phenomena occurring in the system (available in first-principle models of the culture built from process data) in an industrial case study for the development of mAbs in AMBR15® miniature bioreactors. This is achieved by linking the metabolomics data to a kinetic model using data-driven methods to determine the relationships between the most important model parameters for the underlying chemical, physical and biological phenomena, and cell metabolism.

Materials and methods

* + 1. Proposed methodology

In order to relate the information on metabolic dynamics to the macroscopic culture phenomena occurring in CHO cell cultures, a novel methodology is proposed that consists of the following steps (Figure 1):

1. first-principle model building: a state-of-the-art model for CHO cell lines (Kontoravdi et al., 2010) is improved to better describe the contributions of Glutamate and Lactate in order to be representative for the case study considered. The structural identifiability of the model (Villaverde et al., 2016) is verified to determine whether parameter values are unique and meaningful, allowing for reliable model analysis, validation, and prediction;
2. sensitivity analysis: due to limited number of measurements per cell line (namely, $T=7$ time points) of the $V\_{P}$ process variables, a maximum of $P=7$ parameters can be estimated for each cell line. Sensitivity analysis is performed to rank the parameters and identify those which are the most important in describing the specificity of the cell lines behavior. These 7 parameters are referred to as the *most influent* parameters, while the rest of the parameters are referred to as *fixed* parameters;
3. first-principle model parameters estimation: the proposed model is fitted to the process data to estimate the model parameters:
	1. most influent parameter estimation: the set $y$ of the most important parameters is estimated with maximum likelihood estimation (Myung, 2003) from process data for each cell line, while the values of the remaining parameters is fixed;
	2. fixed parameters: cell lines are divided into three groups based on their productivity performances and the fixed parameters are set to a constant value for all the cell lines within a group;
4. machine-learning correlative modelling: the relationships between the cell metabolism dynamics and culture phenomena is studied building a multivariate latent-variable regression model relating the dynamics of the metabolomics data with the most influent parameters;
5. biological understanding: results from machine-learning models are examined to determine which metabolites are the most important in predicting a particular parameter (i.e., highly correlated to a particular phenomenon).



**Figure 1.** *Schematic of the proposed modelling strategy for relating the information on metabolic dynamics to the macroscopic culture phenomena occurring in CHO cell cultures.*

* + 1. Available data and preprocessing

Data from the industrial development of mAb are available for two runs performed in the AMBR15® miniature bioreactor using GSK (Stevenage, UK) proprietary platform process. Both process data and metabolomics data are available.

In particular, a set of $V\_{P}=7 $process variables are measured in $T=7$ time instants during an experimental batch for $N=96$ CHO cell lines: viable cell concentration (VCC), concentration of mAb, Glucose, Glutamate, Glutamine, Lactate and Ammonium. These are arranged in a three-dimensional process dataset $\overline{X}\_{P} [N×V\_{P}×T]$.

For each cell line, intracellular metabolites are analyzed using liquid chromatography-mass spectrometry performed in negative ionization mode. Metabolomics profiles consisting of $V\_{I}=4587$ ions intensities are sampled in two replicates $R=2$ and arranged in a four-dimensional array $\overline{X}\_{I} [N×V\_{I}×(T-1)×R]$. Raw metabolic data is pre-processed for peak detection, global scan alignment, and metabolite annotation (Barberi et al., 2022). Additionally, ions with more than 20% missing intensities are removed from the data set, while missing data imputation (Troyanskaya et al., 2001) is used to impute the remaining missing data. Metabolomics data are then: *i)* variable-wise unfolded in order to account for replications $\overline{X}\_{I}^{'} \left[N∙R ×V\_{I}×\left(T-1\right)\right];$ *ii)* batch-wise unfolded in order to account for data dynamics $X\_{I} \left[N∙R ×V\_{I}∙\left(T-1\right)\right]$; *iii*) Pareto scaled.

* + 1. State-of-the-art kinetic model for CHO cell cultures

The considered state-of-the-art model (Kontoravdi et al., 2010) represents the CHO cell culture behavior through a set of differential equation with 20 parameters and 7 process variables: Volume, Viable Cell Concertation (VCC), Glucose, product titer, Lactate, Glutamine and Ammonia. Unfortunately, this kinetic model does not fully capture the complexity of the system under study and should be improved (see Subsection 3.1).

* + 1. Multivariate statistical analysis

Multiway Partial Least-Square (MPLS) (Nomikos & MacGregor, 1995) consists of a Partial Least-Squares (PLS) model built on a batch-wise unfolded matrix used to consider variables dynamic trajectories. PLS (Wise & Gallagher, 1996) is a linear multivariate regression method which captures the relationship between a matrix of predictors and a matrix of responses. In our case seven models are built, each one using the unfolded metabolomics dataset $X\_{I}$ as a predictor matrix, and a response vector $y=[N×1]$ which collects the values of one of the most influent model parameters for all the cell lines. PLS identifies the direction of maximum covariance between predictors and responses by projecting both $X\_{I} $and $y$ into a reduced space of $A$ Latent Variables (LVs):

|  |  |
| --- | --- |
| $$X\_{I} =TP^{T}+E ,$$$$y=Tq+f ,$$$$T=XW\left(P^{T}W\right)^{-1} ,$$ | (1)(2)(3) |

where $P [A ×V\_{I}]$ and $q [A ×1]$ are the loading matrices, $T [N ×A]$ is the score matrix, $E [N ×V\_{I}]$ and $f [N ×1]$ are the residual matrices of $X$ and $Y$, respectively (minimized in a least-square sense) and $W [N ×A]$ is the weight.

The appropriate number $A$ of LVs is selected by cross-validation (Geladi & Kowalski, 1986). The validation performance of the PLS models is improved through an iterative ion selection procedure retaining only the most informative ions for the prediction (Fernández Pierna et al., 2009).

PLS is used to predict the response $\hat{y}$ (i.e., the relevant model parameter) from a set of new predictors $x\_{NEW} [1×V\_{I}]$:

|  |  |
| --- | --- |
| $\hat{y}=x\_{NEW}B $ | (4) |

where $B=W\left(P^{T}W\right)^{-1}Q^{T}$are the regression coefficients of the PLS model. The PLS model is validated through a Monte-Carlo leave-5-out procedure on $I=10^{5}$ iterations, partitioning the $N$ available samples in 91 cell lines used for calibration and $M=5$ randomly selected for validation. The prediction performance is evaluated through: *i)* determination coefficient in validation $\overbar{Q^{2}}$ averaged throughout the $M$ validation samples and the $I$ iterations; *ii)* index $MMAE/σ$ defined as the median value of$MAE/σ$(mean absolute error of prediction) throughout the $I$ Monte-Carlo iterations, where at each iteration:

|  |  |
| --- | --- |
| $MAE/σ=\frac{\sum\_{m=1}^{M}\left|y\_{m}-\hat{y}\_{m}\right|/M}{σ}$, | (4) |

where $M$ is the number of samples in validation, $y\_{m}$ is the $m$-th sample model parameter, $\hat{y}\_{m}$ is the first-principle model parameter predicted by the PLS model for the same validation samples, and $σ$ is the standard deviation of the considered model parameter used for PLS model calibration.

Results

* + 1. Proposed CHO cell model and structural identifiability

Since the complexity of the system under study is not captured by state-of-art kinetic model of CHO cell cultures (Kontoravdi et al., 2010), the model representativeness is improved for what concerns Glutamate and Lactate consumption (Barberi et al., 2022). In particular, the role of the Glutamate in the culture is added through Equation:

|  |  |
| --- | --- |
| $\frac{dC\_{GLU}}{dt }=\frac{F\_{IN}}{V}\left(C\_{in,glu}-C\_{GLU}\right)-\left(\frac{μ}{Y\_{x,glu}}+m\_{GLU}\right)X\_{V}+\frac{μ}{Y\_{glu,X}}+k\_{1}C\_{GLN}-k\_{2}C\_{GLU}C\_{AMM}$  | (5) |

where $C\_{in,glu}$ is the concentration of Glutamate in a feeding bolus, $k\_{1}$ and $k\_{2}$ are the kinetic constant regulating conversion of Glutamate to Glutamine and $Y\_{gluX}$ and $Y\_{X,glu}$ are the Glutamate production constant. Additionally, Lactate consumption is described a:

|  |  |
| --- | --- |
| $\frac{dC\_{LAT}}{dt}=-\frac{F\_{IN}}{V}C\_{LAT}+Q\_{lat,glc}X\_{V}+\left(Q\_{GLU}Y\_{lat,glu}\right)X\_{V}-\left(\frac{1}{Y\_{x,lat}}\left(\frac{C\_{lat}}{K\_{c,lat}+C\_{lat}}\right)\left(\frac{K\_{c,glc}}{K\_{c,glc}+C\_{GLC}}\right)\right)X\_{V}$  | (6) |

where novel addends regard the last two terms of (6). $Y\_{lat,glu}$ represents the yield of Lactate with respect to Glutamate, $Y\_{xlat}$ is the yield of Lactate consumption, $K\_{C,lat}$ and $K\_{c,glc}$ regulate the Lactate consumption to Glucose. The proposed model is composed of 8 process variables and $25$ parameters, and is structurally identifiable, as verified through the generalized Observability-Identifiability condition (Villaverde et al., 2016).

* + 1. Parameter sensitivity analysis

Parameters sensitivity analysis is performed to identify the parameters that have the largest impact on the first-principle model responses. Sensitivity of each parameter is evaluated by Elementary Effect Test (Saltelli et al., 2008). The most influent parameters that control the phenomena occurring into the culture are: $Y\_{Xglc}$ and $Y\_{Xglu}$ linked to nutrients consumption, $μ\_{max}$ linked to cell growth, $KI\_{amm}$ and $Y\_{latglc}$ related to Ammonia and Lactate inhibition and $Y\_{mAbglc}$ that regulates antibody formation.

The remaining parameters are fixed according to the method described in Subsection 2.1.

* + 1. Biological understanding through PLS modelling

The prediction performance the PLS models predicting the most influential first-principle model parameters from metabolomic dynamics are shown in Table 1 and are reported in term of $\overbar{Q^{2}}$ and$MMAE/σ$. The resulting average determination coefficients for all cell lines range from 45% to 90% showing that a strong relationship between cell metabolism and culture phenomena is found. Additionally, the value of $MMAE/σ$ shows that median absolute errors of prediction are lower than the standard deviation of the parameters used for model calibration.

**Table 1.** *PLS model validation average results for the predicted first-principle kinetic model parameters of all the cell lines.*

|  |  |  |
| --- | --- | --- |
| **Parameter** | **MMAE/σ** | $$\overbar{Q^{2}}$$ |
| $$μ\_{max}$$ | 8.64% | 75.20% |
| $$Y\_{mAbglc}$$ | 7.10% | 50.00% |
| $$Y\_{xglc}$$ | 6.67% | 45.80% |
| $$Y\_{xglu}$$ | 16.30% | 90.40% |
| $$KI\_{amm}$$ | 34.70% | 76.40% |
| $$Y\_{latglc}$$ | 41.30% | 80.90% |
| $$Y\_{glux}$$ | 32.90% | 82.20% |

Furthermore, the PLS models allows the investigation of metabolites associated with the cell culture behaviors. This analysis is performed through the VIP index (Chong & Jun, 2005) and the regression coefficients$B$ of the PLS models. For example, we found that the cell growth and viability is correlated to the metabolism of Arginine and Taurine.

Conclusions

This work proposed a hybrid approach to relate the microscopic information on the metabolism dynamics of the cell lines with the macroscopic chemical, physical and biological phenomena occurring in CHO cell cultures to produce monoclonal antibodies in the biopharmaceutical industry. This was done through first-principle modelling and machine learning. In particular, process data were used to estimate the model parameters of a kinetic model describing the cell behavior. First-principle model parameters, which represent the most important phenomena occurring in the culture, were then linked to metabolic dynamics using data driven approaches. For this purpose, an improved version of a state-of-the-art CHO cell model (Kontoravdi, 2010) was proposed to better represent Glutamate and Lactate consumption. Metabolomics data were then linked to kinetic model parameters using multivariate latent-variable regression models. These highlighted a strong relationship between chemical-physical-biological phenomena and the dynamic evolution of cell metabolism, which allowed understanding the effect of the metabolites on peculiar culture phenomena. For example, cell growth resulted to be related to the metabolism of Arginine and Taurine.

**References**

Ahn, W. S., & Antoniewicz, M. R. (2012). Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnology Journal, 7(1), 61–74.

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Barolo, M., & Facco, P. (2022). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. Metabolic Engineering, 72, 353–364.

Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems, 78(1–2), 103–112.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185(C), 1–17.

Kontoravdi, C., Pistikopoulos, E. N., & Mantalaris, A. (2010). Systematic development of predictive mathematical models for animal cell cultures. Computers & Chemical Engineering, 34(8), 1192–1198.

Kyriakopoulos, S., Ang, K. S., Lakshmanan, M., Huang, Z., Yoon, S., Gunawan, R., & Lee, D. Y. (2018). Kinetic Modeling of Mammalian Cell Culture Bioprocessing: The Quest to Advance Biomanufacturing. Biotechnology Journal, 13(3).

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 47(1), 90–100.

Nomikos, P., & MacGregor, J. F. (1995). Multi-way partial least squares in monitoring batch processes. Chemometrics and Intelligent Laboratory Systems, 30(1), 97–108.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Introduction to Sensitivity Analysis. Global Sensitivity Analysis. The Primer.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. BIOINFORMATICS, 17(6), 520–525.

Villaverde, A. F., Barreiro, A., & Papachristodoulou, A. (2016). Structural Identifiability of Dynamic Systems Biology Models. PLOS Computational Biology, 12(10), e1005153.

Walsh, G., & Walsh, E. (2022). Biopharmaceutical benchmarks 2022. Nature Biotechnology 2022 40:12, 40(12), 1722–1760.

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. Journal of Process Control, 6(6), 329–348.